Methods

A Comparison of Clinically Important Differences in Health-Related Quality of Life for Patients with Chronic Lung Disease, Asthma, or Heart Disease

Kathleen W. Wyrwich, William M. Tierney, Ajit N. Babu, Kurt Kroenke, and Fredric D. Wolinsky

Objective. On the eight scales of the Medical Outcomes Study Short-Form 36-Item Health Survey (SF-36), Version 2, we compared the clinically important difference (CID) thresholds for change over time developed by three separate expert panels of physicians with experience in quality of life assessment among patients with chronic obstructive pulmonary disease (COPD), asthma, and heart disease.
Study Design. We used a modified Delphi technique combined with a face-to-face panel meeting within each disease to organize and conduct the consensus process among the expert panelists, who were familiar with the assessment and evaluations of health-related quality of life (HRQL) measures among patients with the panel-specific disease.
Principal Findings. Each of the expert panels first determined the magnitude of the smallest numerically possible change on each SF-36 scale, referred to as a state change, and then built their CIDs from this metric. All three panels attained consensus on the scale changes that constituted small, moderate, and large clinically important SF-36 change scores. The CIDs established by the heart disease panel were generally greater than the CIDs agreed on by the asthma and COPD panels.
Conclusions. These panel-derived thresholds reflect possible differences in disease management among the represented panel-specific diseases, and are all greater than the minimal CID thresholds previously developed for the SF-36 scales among patients with arthritis. If confirmed among patients with the relevant diseases and those patients' physicians, these disease-specific CIDs could assist both researchers and practicing clinicians in the use and interpretation of HRQL changes over time.

Key Words. Quality of life, chronic obstructive pulmonary disease, asthma, coronary artery disease, congestive heart failure, consensus panel

The incorporation of patient-reported health-related quality of life (HRQL) measures to better assess clinical outcomes has been an important goal of evidence-based medicine (Guyatt et al. 1997). The Medical Outcomes Study Short-Form 36-Item Health Survey, or SF-36, is the most widely used HRQL instrument in the world (Brazier, Harper, and Jones 1992). Designed as a generic instrument capable of measuring the HRQL of individuals with different diseases or health conditions, the SF-36 yields scale scores for eight domains: Physical Functioning, Role Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role Emotional, and Mental Health. Recently, this instrument was revised to reflect improved wording and response options, and Version 2 is now available for patient-reported measurement of HRQL outcomes (Ware, Kosinski, and Dewey 2000).

Other HRQL measures, primarily disease-specific instruments designed to measure particular health conditions, have established standards for determining clinically important differences (CIDs) using various approaches for interpreting important changes over time (Guyatt et al. 2002). Many consumers seek such standards or thresholds for interpreting and evaluating change on these important outcomes (Symonds et al. 2002). These consumers include not only patients, clinicians, and clinical researchers but also pharmaceutical and medical device manufacturers who must demonstrate the usefulness of their products, and government regulators who, along with insurance payers, must evaluate the usefulness and consequences of each product seeking endorsement or coverage. Without established standards for interpreting the change in HRQL measures attributed to treatments or interventions, these consumers must often resort to statistical evaluations that rely on the variation in a sample(s) and the number of enrollees or power to detect a statistically significant difference ($p < .05$) between two groups, such as treatment versus placebo. Statistically significant differences, however, do not imply that a meaningful or relevant difference has been demonstrated for the individuals enrolled in such trials (Sloan et al. 2002).

Despite the popularity of the SF-36 as a clinical and research tool, there had been no studies investigating standards for determining CIDs in SF-36 scale scores for individual patients until quite recently (Kosinski et al. 2000). Kosinski and colleagues examined change in SF-36, Version 1 scales and

Address correspondence to Kathleen W. Wyrwich, Ph.D., Departments of Research Methodology and Health Services Research, Saint Louis University, 221 N. Grand Avenue, St. Louis, MO, 63103. William M. Tierney, M.D., is with the Division of General Internal Medicine and Geriatrics, Indiana University School of Medicine, Wishard Memorial Hospital, Indianapolis, IN. Kurt Kroenke, M.D., is with the Regenstrief Institute and Indiana University School of Medicine, Indianapolis, IN. Ajit N. Babu, M.B.B.S., M.P.H., is Professor and Chairman, Institute for Medical Informatics and Multimedia Education, Amrita Institute of Medical Sciences, Cochin, Kerala, India. Fredric D. Wolinsky, Ph.D., is with the Department of Health Management and Policy, College of Public Health, The University of Iowa and with the General Hospital, Iowa City, IA.

the Health Assessment Questionnaire (HAQ) among patients with rheumatoid arthritis (RA) and their physicians. Using several anchors commonly used to measure change in RA severity that included: (1) a 1–19 percent improvement in the number of swollen joints; (2) a 1–19 percent improvement in the number of tender joints; (3) a patient global pain assessment; (4) a patient global overall change assessment; and (5) a physician global change assessment, Kosinski et al. calculated different minimal CID levels for each of these anchors by averaging the mean change scores across the patients displaying one level of improvement. These results, however, varied widely across the five criteria for important change. The authors eventually concluded that:

> Although this study did not establish the single best estimate of a minimally important change in the SF-36 and HAQ scores, the results establish a range of estimates across the various dimensions measured by the HQL [HRQL] instruments within which a minimally important change probably occurs. (p. 1486)

Recognizing the need for interpretation standards that clinicians and other stake holders can easily use to benchmark important changes in the HRQL, we began addressing the need for CIDs for the eight scales of the SF-36, Version 2, by assembling three expert panels of North American physicians to identify HRQL change thresholds for the SF-36 in patients with chronic obstructive pulmonary disease (COPD), asthma, or heart disease (coronary artery disease and congestive heart failure). These three disease areas represent common chronic conditions among patients treated in primary care settings, and exemplify the noteworthy challenge of short- and long-term assessment of HRQL among chronic disease patients beyond the clinical markers, such as spirometry or echocardiogram results. This report compares and further examines the CID results from each of these independent panels for the SF-36, Version 2 scales. By comparing the panels' results across the domains of the SF-36, we can further understand commonalities and disease-specific differences in how the expert physician panels viewed CIDs in HRQL among patients within each of these three diseases. In addition, a deeper examination of these SF-36 panel reports allow for further comparisons of this method with other approaches to develop important change thresholds for the scales of this instrument.

METHODS

We utilized elements of the Delphi technique and borrowed components of the RAND/UCLA Appropriateness Method to assemble and direct the consensus process for each of the three panels (Brook et al. 1986; McGlynn, Kosecoff, and Brook 1990). We began by conducting several searches of the *Medline* database to find North American physician authors publishing longitudinal studies of HRQL among patients with COPD, asthma, or heart disease using the HRQL instruments of interest (SF-36, the Chronic Respiratory Disease Questionnaire [CRQ] [Guyatt 1988], the Asthma Quality of Life Questionnaire [AQLQ] [Juniper et al. 1993], or the modified version of the Chronic Heart Failure Questionnaire [CHQ] [Wolinsky et al. 1998]). These instruments were chosen over other generic and disease-specific HRQL measures because of their previously demonstrated psychometric qualities, as well as their relatively straightforward use and scoring in both research and clinical settings (Wolinsky et al. 1998; Wyrwich, Nienaber et al. 1999; Wyrwich, Tierney, and Wolinsky 1999, 2002). We evaluated the physicians on these three disease-related author lists (COPD, asthma, or heart disease) for appropriateness to insure that by the nature of their clinical and research accomplishments, each potential panelist was a principal in their field's measurement of HRQL. All were considered as suitable. We then queried the physician authors about their interest in serving on a panel and their availability. From those physicians who responded to our queries, we selected nine panelists for each of the three target disease areas who represented a balance between specialists and generalists, geographic diversity, and other relevant factors. We also selected a panel chairperson from the nine panelists for each targeted disease. The respective panels and each panelist's home institutions are listed at the end of this report.

To begin developing CIDs for the SF-36 scales, each panel completed two Delphi rounds. First, the panel was sent peer-reviewed articles that used or evaluated the SF-36 among patients with their target disease. They were instructed to read each article and then estimate CIDs for each scale of the SF-36, as well as the appropriate disease-specific HRQL measure (CRQ, AQLQ, or CHQ). It is important to note that panelists were not provided with any specific definition of a CID, but left to determine their own meaning for this term that was reflected in their choice for small, moderate, and large change levels for improvements or declines. The first round CID estimates were compiled anonymously and sent back to the panelists within each target disease in Round 2, along with additional relevant literature suggested by the respective panelists. Each panel then provided revised CIDs estimates that we again anonymously compiled and sent back to them to peruse before each panel met in person.

Three panel meetings, one for each target disease, took place in St. Louis, Missouri, during May and June of 2000. The panel meetings lasted approximately 4 hours. An a priori agreement stated that if no more than two of the reviewers disagreed, the majority opinion would be considered a consensus. Each panel reached consensus on all SF-36 scales. After each panel meeting, the respective panel chairperson prepared a report of the panel's work, which was iteratively circulated and modified by the panel members until all agreed that it reflected an accurate account of their meeting.

RESULTS

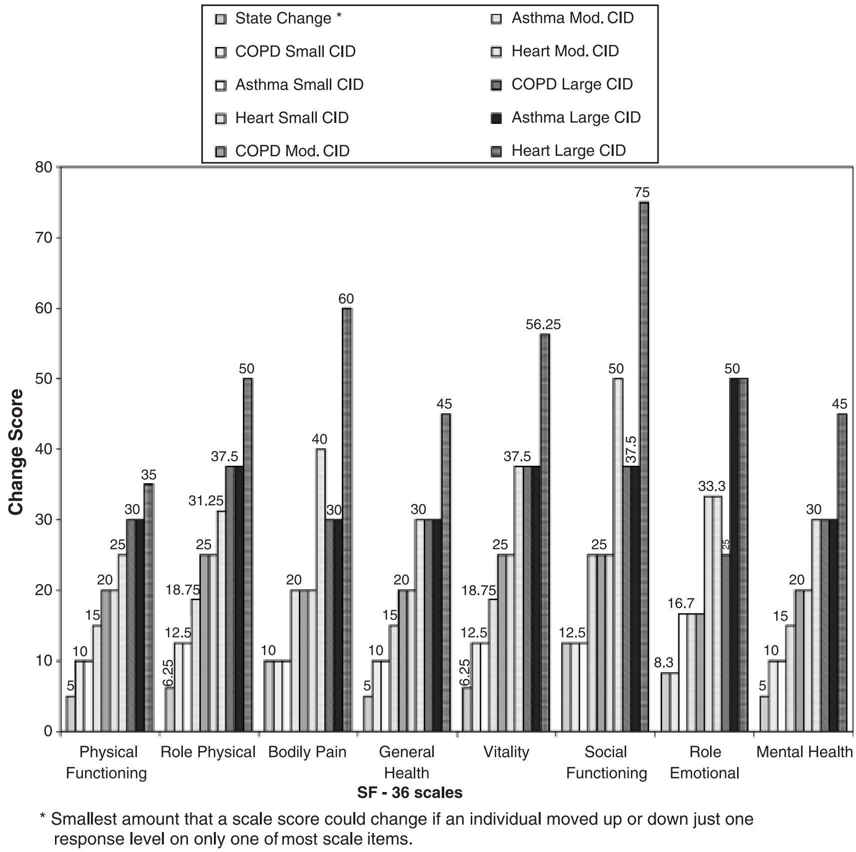
In developing their CIDs, each panel meeting began with a review of the Delphi Round 2 results, and then an individual statement by each panelist regarding how s/he went about determining the CIDs thresholds. The methods for shaping panelist's individual assessments included: (1) experience with patient HRQL change data in clinical trials and practice; (2) exploration of the percent of change (20, 40 percent, etc.) associated with the base scoring range; (3) triangulation among results from CID studies using the same HRQL measures; (4) the use of Cohen's effect size thresholds (small = 0.2, moderate = 0.5, large = 0.8) (Cohen 1977); (5) the examination of individual patient change scenarios for changes in clinical conditions and associated changes on the SF-36 scales (6) a review of the literature from recognized pharmaceutical treatments where HRQL score changes were reported; (7) the application of prospect theory, which asserts that healthier patients value small declines to be of greater importance than sicker patients (Kahneman and Tversky 1979); and (8) the comparison of SF-36 scores among known groups (inpatients and outpatients) with the targeted condition.

After each panelist provided their views, the panelists at each meeting then derived the smallest amount that a scale score could change if an individual patient improved or worsened by one response level on just one of most scale items, and labeled this amount as a *state change.* For example, there are ten items in the Physical Functioning scale, and each item has three response choices: not limited at all, limited a little, limited a lot. According to the developer's scoring instructions, each SF-36 scale score is transformed to a 0–100 scale (worst–best) (Ware, Kosinski, and Dewey 2000). If between two separate clinician office visits a patient improved only his/her ability to climb a flight of stairs (change from limited a little to not limited at all), and all responses to the other nine Physical Functioning items remained the same (unchanged), the patient's scale score at the second office visit would increase by one state change, or in the case of the Physical Functioning scale, 5 points.

Because the number of items and number of response choices vary across the eight SF-36 scales, panelists carefully scrutinized each 0–100 scale to determine the value of one state change. The magnitude of the smallest possible change is given in the first bar for each SF-36 scale in Figure 1.

Once the value of a single state change had been calculated, the panelists then discussed how many state changes would be needed by a patient in their disease area in order for the change to be considered small, but clinically important. Thus, all of the panels' CIDs are multiples of the specific SF-36

Figure 1: Clinically Important Difference Levels for Change over Time on the Eight Scales of the Medical Outcomes Study Short-Form 36-Item Health Survey, Version 2, Established by Three Expert Panels of Physicians (Chronic Obstructive Pulmonary Disease [COPD], Asthma, and Heart Disease)



scale's state change value. For example, the Asthma panel arrived at a consensus asserting that a small CID on the Physical Functioning Scale is 10 points, or the equivalent of 2 net state changes across the scale's 10 items. After deciding on the magnitude of a small CID, the panels then considered appropriate levels for moderate and large CIDs for each SF-36 scale. Figure 1 displays the small, moderate, and large CIDs established by the three disease-specific panels for the eight scales of the SF-36, Version 2.

In all three panel discussions, panelists made a deliberate decision for the sake of simplicity to equate improvements and declines at each level. In other words, the absolute value of a small but clinically important decline would equal that of a small but clinically important improvement. Each panel also discussed the issues surrounding patients with differing disease severity and comorbidities, and then gauged their results on typical presenting patients with their target disease.

With one scale's exception, the resulting small, moderate, and large CIDs are the same for the COPD and the asthma panels. The heart disease CIDs, however, tend to be larger than those of asthma and COPD. The Role Emotional scale is the exception to this pattern, where the asthma panel's CID levels equaled the heart disease panel's estimate at all CID levels.

For all SF-36 scales, the COPD and asthma moderate CID levels are twice as large as the respective panels' levels for a small CID. Likewise, the large CID levels for the COPD and asthma panels are three times the magnitude of their small CID levels. The heart disease panel's moderate and large CIDs are also two and three times the small CID estimates for all SF-36 scales—with the exception of the Physical Functioning and Role Physical scales. On these scales, the leaps from small to moderate and moderate to large each require two additional state changes to occur.

DISCUSSION

Using a modified Delphi method, three expert panels of physicians with experience using HRQL measures among patients with COPD, asthma, or heart disease reached consensus on the magnitude of change needed to achieve small, moderate, and large CIDs for each scale of the SF-36, Version 2. The panels of experienced research physicians built their results by extracting the minimal amount that each scale's score can numerically change (a state change) and then determining how many state changes are needed in order for the change to be considered clinically important.

In general, the CIDs for COPD and asthma were quite similar, while the CIDs for heart disease were higher. A plausible explanation for the generally higher CIDs established by the panel of expert physicians in heart disease may come from deliberations during their panel meeting over what portion of cardiac patients' SF-36 change scores are not the result of clinically important changes in their heart disease (Wyrwich, Spertus et al. 2004). This led to a discussion among those panelists about the appropriateness of using a generic instrument with heart disease patients, and the panelists' general preference for disease-specific measures that better reveal the improvements in their patients' HRQL that proper cardiac treatment can provide. Nonetheless, the heart disease panel continued to use higher CIDs when they established change standards for their disease-specific instrument, the CHQ. Two domains of the CHQ, dyspnea and fatigue, mirror the items on the CRQ, the disease-specific instrument examined by the COPD panel. Consistent with the SF-36 comparisons, the heart disease panel's CIDs for the CHQ were also higher than the COPD panel's CIDs for the CRQ in these two domains (Wyrwich, Fihn et al. 2003; Wyrwich, Spertus 2004).

Another possible explanation for the comparability of CIDs in COPD and asthma in contrast to the heart disease panel results could be the similarity of disease process for the two obstructive airway conditions (Wyrwich, Fihn et al. 2003; Wyrwich, Nelson et al. 2003). These diseases, which may be more sensitive to fluctuations in the environment, may also lead to greater independence among COPD and asthma patients in the management of minor exacerbations (Juniper 2003). Thus, the obstructed airway disease patients may be more resistant in reporting any changes in their HRQL to their physician unless the differences are of greater individual importance, resulting in a more sensitive calibration of change detection among physicians treating patients with obstructive airway disease.

The work of these panels in determining CIDs sheds light on evaluating SF-36 change scores from other patient samples and general population samples. As shown in Figure 1, the minimal amount of change that is numerically possible for an SF-36 scale is at least 5 points, and ranges up to 12.5 points on the Social Functioning scale. These findings we expect will discourage others in future interpretations of SF-36 change scores from relying on the often-cited standard that a 3–5 point shift represents an important difference on any SF-36 scale. This standard was derived from a study using the short-lived SF-20 HRQL instrument (Stewart, Greenfield, and Hays 1989), and based the minimal difference magnitude on a statistically significant cross-sectional difference between two very large-sized groups. As we previously stated, although a statistically significant difference for a given p-value is achieved, this difference may not represent a clinically relevant differentiation and needs to be investigated at a more meaningful individual level.

As stated in the Introduction of this report, only one other known study (Kosinski et al. 2000) has attempted to find small but clinically important change levels for the eight scales of the SF-36. Most of the mean change results from the five criteria that anchored improvement scores in that study are smaller than one state change on each of the eight SF-36 scales, and only the Physical Functioning, Bodily Pain, Vitality, and Mental Health domains had any mean change estimates that equaled at least one state change. For example, the mean change scores for the Role Physical, Social Functioning, or Role Emotional scales did not change as much as one state change for any of the five RA severity criteria. Moreover, the mean changes on each SF-36 scale anchored on the physician's global assessments of each patient's overall improvement by one level (e.g., fair to good) were marginally lower in all domains than the mean change scores driven by patient-reported global overall change assessments for one level of improvement (Kosinski et al. 2000).

In contrast, the panels' consensus change assessments are much larger in all three targeted diseases than the physician- and patient-specific criteria assessment conducted by Kosinski et al. (2000) among RA patients. Focusing on the value of a state change on each SF-36 scale illuminates the individual differences that the expert panelists viewed as important improvements or declines. The interpretation and translation of observational and intervention research using HRQL measures as outcomes can be directly improved if authors provided not only group mean changes and standard deviations but also a reporting of the proportion of patients who improved/declined by a small, moderate, or large amounts using CIDs (Guyatt et al. 2002). The large differences between our panels' CIDs and those from the RA patients, however, may lead consumers of these standards to question whether these small but clinically important change estimates are within the limits of patient measurement error on the SF-36, and, therefore, relatively meaningless in the evaluation of patient HRQL change.

We can begin to tackle this question by examining the standard error of measurement (SEM). If a patient was repeatedly measured many times on a measure and s/he did not change during these repeated measurements, nor remember the prior scores (independent assessments), a frequency graph of these scores should result in a normal distribution centered on the patient's true scores. The standard deviation of that normal distribution is the SEM. Specifically, this statistic is calculated using the reliability of a scale (r_{tt}) and its standard deviation (σ) or

$$\mathrm{SEM} = \sigma\sqrt{1 - r_{tt}}$$

Thus, the 95 percent confidence interval around an individual patient's scores is calculated using $\pm$ 1.96 SEM. Indeed, the current method for classifying HRQL change (improved, stable, or declines) in health status over 2-year intervals among enrollees in Medicare managed care plans in the sizeable Health Outcomes Survey cohorts evaluates individual shifts over time on the SF-36 Physical Component Scores (PCS) and Mental Component Scores (MCS) using an increase of at least 1.96 SEM as the threshold for improvement and a decline of at least −1.96 SEM to define getting worse (Centers for Medicare and Medicaid Services 2003). These same $\pm$ 1.96 SEM thresholds were also used in the 1996 report classifying HRQL changes over a 4-year time span among the Medical Outcome Study enrollees, again based on PCS and MCS change scores (Ware et al. 1996).

We, however, did not ask the physician panels to estimate the CID thresholds for the PCS and MCS during the consensus process because of the computation and conceptual difficulties associated with these estimates. Basically, the calculation of the PCS and MCS incorporates factor score coefficients (from a two-factor solution factor analysis among the eight SF-36 scale scores) and standardized SF-36 scale scores, and then transformations of the resulting calculations to a norm-based score with a mean of 50 and a SD of 10 (Ware, Kosinski, and Dewey 2000). And even with the most extensive physician–researcher knowledge of the SF-36 used among disease-specific patients, such conceptualization for identifying important change scores in individual patients would have been an extremely challenging, if not daunting, task.

The 95 percent confidence intervals ($\pm$ 1.96 SEM) around eight scale scores of the SF-36, Version 2, range in width from 12 to 17 points, based on a 1998 general U.S. population sample (Ware, Kosinski, and Dewey 2000). This range is as large, if not larger, than nearly all of the physician panels' estimates for small CIDs, and is also larger than the extra magnitude that the heart disease panel's results have over COPD and asthma small CID consensus estimates. Others, however, have demonstrated that just 1 SEM corresponds to the minimal CID threshold for several disease-specific HRQL measures among chronic disease patients (Wyrwich, Nienaber et al. 1999; Wyrwich, Tierney, and Wolinsky 1999, 2002; Cella et al. 2002). Although we do not have a 95 percent level of confidence that this difference is outside the limits of measurement error, the 1 SEM threshold is well beyond the necessary "more than likely" or 51 percent level of confidence that change occurred at the individual level (Donaldson and Moinpour 2002; Wyrwich 2004). At the 1 SEM level (6–8.5 points on the SF-36, Version 2 scales), the small CIDs reported by our panelists are beyond this lower standard for individual measurement error, and most differences between the heart panel results compared with the COPD and asthma panels are relevant. However, several of the SF-36 scale estimates for a minimal clinically important change among RA patients remain below the 1 SEM threshold, as well as below the magnitude of one state change for their respective scale.

There are several significant limitations to this comparison paper that speak primarily to the expert panel consensus process. It is possible that unknown to us, panel dynamics led to bias, pressure, or intimidation (Stasser, Kerr, and Davis 1989). We believe, however, that our consensus procedures, which allow discordance to be voiced, avoided any coercion. In addition, the Delphi process has been used in numerous settings to "allow a reasonable consensus to be developed by a group of experienced individuals" (Bellamy et al. 1991, p. 1719). Other key limitations include the use of only a single panel for each targeted disease (Shekelle et al. 1998), and the assumption of symmetry between the magnitude of improvements and declines.

An equally important limitation to this paper is the "easy road" that we have initially taken as we explore the meaning and magnitude of a clinically important change in HRQL. By assembling physician researchers with experience among the same patient group, we were able to seek a consensus opinion on the reported CIDs, and because of their experience with both the treatment of patients with the targeted diseases and with the evaluation instruments, these expert physicians' input provides some clinical evidence for invoking the *clinically important difference* label. However, the measures we are investigating are, indeed, patient-reported, and patient-informed thresholds for important change are needed, as well as a more directly informed method to incorporate a clinical perspective. Undeniably, in an ideal approach to medical decision making, clinicians and other consumers of CID standards would always use and incorporate their knowledge of the patient's preferences and values. Therefore, it is important to articulate that we have chosen not to value the panels' opinions as "greater" than patients', despite the focus of this comparison report. Instead, in addition to the expert panel data, we are also gathering two other streams of data as part of a large multicentered study. One stream involves nearly 1,700 patients and their HRQL scores, as well as patient-perceived changes in both disease-specific and generic HRQL domains, measured every 2 months across a year of enrollment. A second stream of data comes from these patients' primary care physicians, who report on their patients' HRQL at enrollment and at subsequent office visits during the year of follow-up. Ultimately, triangulating our expert consensus results with the perspectives of patients and their treating physicians will provide the fullest perspective of how to best define and compare CIDs on HRQL measures. If the CID thresholds established by our expert panels are confirmed in studies among patients with these target diseases and their treating clinicians, the results would facilitate the interpretation of HRQL changes over time among patients with COPD, asthma, or heart disease. This could lead to HRQL scores being more clinically useful to clinicians. Moreover, the interpretation and translation of research using HRQL measures as outcomes could be more relevant to patients, their healthcare providers, and other consumers of important change standards.

ACKNOWLEDGMENTS

Results in this manuscript were presented at The International Society for Quality of Life Research 8th Annual Meeting on November 9, 2001 in Amsterdam, the Netherlands and at the Academy of Health Services Research and Health Policy Annual Meeting on June 24, 2002. This research was funded by grants from the Agency for Healthcare Research and Quality to Dr. Wolinsky (R01 HS10234) and Dr. Wyrwich (K02 HS11635). The authors thank Janet Bafia and Sharon Fryer for transcribing the audio- and videotapes from the Panel Meeting, Stacie Metz for her work in the preparation of this manuscript, and Kelli Norton and Joe Kesterson for coordinating the data and chart searches required to create the Round 1 materials. The authors also acknowledge the valuable assistance of the Expert Panel Members:

COPD Expert Panel	Heart Disease Expert Panel	Asthma Expert Panel
Stephan D. Fihn, M.D. Panel Chair *University of Washington*	John A. Spertus, M.D. Panel Chair *University of Missouri*	Harold S. Nelson, M.D. Panel Chair *National Jewish Medical & Research Center*
Robert A. Barbee, M.D. *University of Arizona*	Kirkwood F. Adams, M.D. *University of N. Carolina*	Andrea J. Apter, M.D. *University of Pennsylvania*

(*Continued*)

COPD Expert Panel	Heart Disease Expert Panel	Asthma Expert Panel
James F. Donohue, M.D. *University of N. Carolina*	Ronald S. Baigrie, M.D. *Sudbury Regional Hospital*	Jonathan A. Bernstein, M.D. *Bernstein Clinical Research Center*
Nicholas J. Gross, M.D. *Hines VA Hospital*	Marshall H. Chin, M.D. *University of Chicago*	Robert A. Nathan, M.D. *Allergy & Asthma Associates*
Richard V. Hodder, M.D. *University of Ottawa*	Donald J. Mertens, M.D. *Willowdale, Ontario*	Rebecca R. Roberts, M.D. *Cook County Hospital*
Donald A. Redelmeier, M.D. *University of Toronto*	Michael W. Rich, M.D. *Washington University*	D. Robert Webb, M.D. *Allergy Associates, PC*
Kathleen A. Rickard, M.D. *Glaxo Wellcome, Inc.*	Kenneth Rockwood, M.D. *Dalhousie University*	Bennett P. DeBoisblanc, M.D. *Louisiana State University*
E. P. Trulock, M.D. *Washington University*	Roy J. Shephard, M.D. *Brackendale, British Columbia*	Leonard Bielory, M.D. *New Jersey Medical School*
Roger D. Yusen, M.D. *Washington University*	Robert J. Zalenski, M.D. *Wayne State University*	

REFERENCES

Bellamy, N., W. W. Buchanan, J. M. Esdaile, A. G. Fam, W. E. Kean, J. M. Thompson, G. A. Wells, and J. Campbell. 1991. "Ankylosing Spondylitis Antirheumatic Trials. III. Setting the Delta for Clinical Trials of Anti-Rheumatic Drugs–Results of a Consensus Development (Delphi) Exercise." *Journal of Rheumatology* 8: 1716–22.

Brazier, J., R. Harper, and N. Jones. 1992. "Validating the SF-36 Health Survey Questionnaire: New Outcome Measure for Primary Care." *British Medical Journal* 305: 160–4.

Brook, R., M. Chassin, A. Fink, D. Solomon, J. Kosecoff, and R. Park. 1986. "A Method for the Detailed Assessment of the Appropriateness of Medical Technologies." *International Journal of Technology Assessment in Health Care* 2: 53–63.

Centers for Medicare and Medicaid Services. 2003. "Medicare Health Outcomes Survey *Cohort III* Performance Measurement Report" [accessed on June 21, 2004]. Available at http://www.cms.hhs.gov/surveys/hos/download/HOS_Sample-PMR_C3.pdf.

Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press.

Donaldson, G., and C. Moinpour. 2002. "Individual Differences in Quality-of-Life Treatment Response." *Medical Care* 40 (6 suppl): III39–53.

Guyatt, G. 1988. "Measuring Health Status in Chronic Airflow Limitation." *European Respiratory Journal* 1: 560–4.

Guyatt, G. H., C. D. Naylor, E. Juniper, D. Heyland, R. Jaeschke, and D. Cook. 1997. "Users' Guides to the Medical Literature. XII. How to Use Articles about Health-Related Quality of Life Measurements." *Journal of the American Medical Association* 277 (15): 1232–7.

Guyatt, G., D. Osoba, A. Wu, K. Wyrwich, G. Norman, and the Clinical Significance Consensus Meeting Group. 2002. "Methods to Explain the Clinical Significance of Health Status Measures." *Mayo Clinic Proceedings* 77: 371–83.

Juniper, E. 2003. "Interpreting Quality of Life Data: Should We Listen to the Patient or the Clinician?" *Annals of Allergy, Asthma and Immunology* 91 (2): 115–6.

Juniper, E., G. Guyatt, P. Ferrie, and L. Griffith. 1993. "Measuring Quality of Life in Asthma." *American Review of Respiratory Diseases* 127: 832–8.

Kahneman, D., and A. Tversky. 1979. "Prospect Theory: An Analysis of Choice under Risk." *Econometrica* 47 (2): 263–91.

Kosinski, M., S. Z. Zhao, S. Dedhiya, J. T. Osterhaus, and J. E. Ware Jr. 2000. "Determining Minimally Important Changes in Generic and Disease-Specific Health-Related Quality of Life Questionnaires in Clinical Trials of Rheumatoid Arthritis." *Arthritis and Rheumatism* 43: 1478–87.

McGlynn, E., J. Kosecoff, and R. Brook. 1990. "Format and Conduct of Consensus Development Conferences: Multi-Nation Comparison." *International Journal of Technology Assessment in Health Care* 6 (3): 450–69.

Shekelle, P., J. Kahan, S. Bernstein, L. Leape, C. Kamberg, and R. E. Park. 1998. "The Reproducibility of a Method to Identify the Overuse and Underuse of Medical Procedures." *New England Journal of Medicine* 338: 1888–95.

Sloan, J. A., D. Cella, M. H. Frost, G. H. Guyatt, M. A. G. Sprangers, T. Symonds, and the Clinical Significance Consensus Meeting Group. 2002. "Assessing Clinical Significance in Measuring Oncology Patient Quality of Life: Introduction to the Symposium, Content Overview, and Definition of Terms." *Mayo Clinic Proceedings* 77: 367–70.

Stasser, G., N. Kerr, and J. H. Davis. 1989. "Influence Processes and Consensus Models in Decision-Making Groups." In *Psychology of Group Influence*, 2d ed, edited by P. Paulus. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Stewart, A., S. Greenfield, and R. Hays. 1989. "Functional Status and Well-Being of Patients with Chronic Medical Conditions: Results from the Medical Outcomes Study." *Journal of American Medical Association* 262: 907–13.

Symonds, T., R. Berzon, P. Marquis, T. Rummans, and the Clinical Significance Consensus Meeting Group. 2002. "The Clinical Significance of Quality-of-Life

Results: Practical Considerations for Specific Audiences." *Mayo Clinic Proceedings* 77: 572–83.

Ware, J. E. Jr., M. S. Bayliss, W. H. Rogers, M. Kosinski, and A. R. Tarlov. 1996. "Differences in 4-year Health Outcomes for Elderly and Poor, Chronically Ill Patients Treated in HMO and Fee-for-Service Systems. Results from the Medical Outcomes Study." *Journal of American Medical Association* 276 (13): 1039–47.

Ware, J., M. Kosinski, and J. Dewey. 2000. *How to Score Version Two of the SF-36 Health Survey.* Lincoln, RI: QualityMetric Inc.

Wolinsky, F., K. Wyrwich, N. Nienaber, and W. Tierney. 1998. "Generic vs. Disease-Specific Health Status Measures: An Example Using Coronary Artery Disease and/or Congestive Heart Failure Patients." *Evaluation and the Health Professions* 21 (2): 216–43.

Wyrwich, K. 2004. "Minimal Important Difference Thresholds and the Standard Error of Measurement: Is There a Connection?" *Journal of Biopharmaceutical Statistics* 14 (1): 97–110.

Wyrwich, K., S. Fihn, W. Tierney, K. Kroenke, A. Babu, and F. Wolinsky. 2003. "Clinically Important Differences in Health-Related Quality of Life for Patients with Chronic Obstructive Pulmonary Disease: An Expert Panel Report." *Journal of General Internal Medicine* 18 (3): 196–202.

Wyrwich, K., H. Nelson, W. Tierney, K. Kroenke, A. Babu, and F. Wolinsky. 2003. "Clinically Important Differences in Health-Related Quality of Life for Patients with Asthma: An Expert Panel Report." *Journal of Asthma, Allergy and Immunology* 91 (2): 148–53.

Wyrwich, K., N. Nienaber, W. Tierney, and F. Wolinsky. 1999. "Linking Clinical Relevance and Statistical Significance in Evaluating Intra-Individual Changes in Health-Related Quality of Life." *Medical Care* 37 (4): 469–78.

Wyrwich, K., J. Spertus, K. Kroenke, W. Tierney, A. Babu, and F. Wolinsky. 2004. "Clinically Important Differences in Health Status of Life for Patients with Heart Disease: An Expert Panel Report." *American Heart Journal* 147 (4): 615–22.

Wyrwich, K., W. Tierney, and F. Wolinsky. 1999. "Further Evidence Supporting a SEM-Based Criterion for Identifying Meaningful Intra-Individual Changes in Health-Related Quality of Life." *Journal of Clinical Epidemiology* 52 (9): 861–73.

———. 2002. "Using the Standard Error of Measurement to Identify Important Intra-Individual Change on the Asthma Quality of Life Questionnaire." *Quality of Life Research* 11 (1): 1–7.